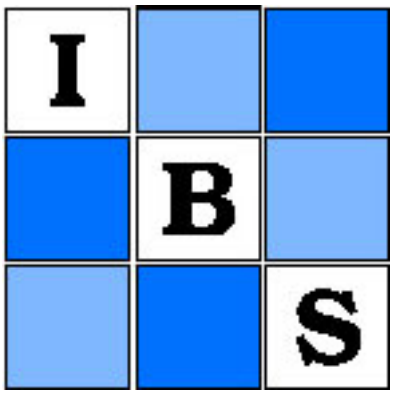

BIOMETRICS 44, 35–47
March 1988

Regression Trees for Censored Data

Mark Robert Segal
Channing Laboratory, Harvard Medical School,
180 Longwood Avenue, Boston, Massachusetts 02115, U.S.A.

SUMMARY

The regression-tree methodology is extended to right-censored response variables by replacing the conventional splitting rules with rules based on the Tarone–Ware or Harrington–Fleming classes of two-sample statistics. New pruning strategies for determining desirable tree size are also devised. Properties of this approach are developed and comparisons with existing procedures, in terms of practical problems, are discussed. Illustrative, real-world performances of the technique are presented.

1. Introduction

In recent years considerable research effort has been dedicated toward devising regression techniques free from some of the restrictive classical assumptions. Frequently, the tradeoff for this freedom is that these methods require large data sets and are computationally intensive. Two broad categories of such approaches can be summarized as *smoothing* (Stone, 1977; Friedman and Stuetzle, 1981; Breiman and Friedman, 1985) and *trees* (Breiman et al., 1984). Another area of current statistical endeavor is the analysis of survival data wherein the response variable is subject to censoring. The proportional hazards model (Cox, 1972) has afforded a widely used and flexible technique. However, again constraining assumptions have become an issue and attempts to overcome these have led to the adaption of smoothing and tree methodologies to the survival setting. Specifically, Hastie and Tibshirani (1986) extend the Cox model by replacing the linear modeling of covariates with an additive (sum of smooth functions) model.

This paper proposes modifications to the conventional regression tree methodology with the primary motivation of facilitating an extension to censored data. However, the suggested changes used to achieve this end have merit in their own right, and so comparisons with the existing techniques (in the uncensored setting) are also mentioned. The basic alteration is the replacement of the goodness-of-split criteria, which had been geared toward optimizing within-node *homogeneity*, with measures of between-node *separation*. The measures used are two-sample statistics belonging to the Tarone–Ware or Harrington–Fleming classes. Their introduction further necessitates changing the pruning algorithm used to determine desirable tree size.

Existing methods for tree construction in the survival data setting utilize different splitting and pruning approaches. Gordon and Olshen (1985) use distance measures (Wasserstein metrics) between Kaplan–Meier curves and certain point masses. The motivation is by way of analogy with the least squares criterion used in the uncensored setting and the method allows for an immediate inheritance of the *CART* (see below) pruning algorithm. Ciampi et al. (1987) and Davis (unpublished Ph.D. thesis, Harvard University, 1988) pursue splitting based on likelihood-ratio tests. Work on comparing the performance and properties of the various tree techniques is in progress.

Key words: Censoring; Pruning; Regression tree; Splitting rule; Tarone–Ware class.

The next section provides a brief overview of current regression tree (or recursive partitioning) methodology. Section 3 deals with the new splitting criteria and, in particular, addresses the censored data issue. Section 4 indicates how the new pruning strategies work. The fifth section demonstrates the performance of the new technique on a real-world example, and the sixth discusses properties and potential of the new approach with reference to problems encountered in practice.

Repeated allusion is made to the definitive reference by Breiman et al. (1984), which is referred to as *CART*, and in which stand-alone section numbers should be sought. This monograph makes explicit the advantages of tree methods from an applied perspective. The importance of tree techniques in biomedical settings has also been emphasized by Goldman et al. (1982). The extraction of clinically meaningful strata that have distinct survival prospects is an endpoint commonly sought by medical investigators. Further, the tree structure makes for ready interpretability and easy classification of new patients.

2. Constructing Regression Trees

A simplified description of regression trees is presented in this section, so that the subsequent reformulations can be understood. For a more detailed understanding the *CART* monograph is recommended. Attention here is restricted to the familiar regression setting—there are p predictor variables $X_1, \ldots, X_p$ and a continuous and uncensored response Y. No comment is made here with regard to issues such as the treatment of missing values (§5.3.2) or variable importance (§5.3.4), for which carryover from the standard methods to the new methods is immediate. For a compilation of the advantages afforded by using tree procedures see Segal (unpublished Ph.D. thesis, Department of Statistics, Stanford University, 1986) or *CART* §2.7.

In order to construct a regression tree four components are required. These are:

1. A set of (binary) questions of the form "Is $\vec{x} \in A$?" where $\vec{x}$ is a case and $A \subset \mathscr{X}$, the predictor space. The answer to such a question induces a partition, or split, of the predictor space. That is, the cases for which the answer is *yes* are associated with region A and those for which the answer is *no* are associated with the complement of A. The subsamples so formed are called *nodes*.
2. A goodness-of-split criterion $\phi(s, t)$ that can be evaluated for any split s of any node t. The criterion is used to assess the worth of the competing splits, where (in *CART*) worth pertains to within-node homogeneity.
3. A means for determining the appropriate tree size.
4. Statistical summaries for the terminal nodes of the selected tree. These can be as simple as the node average or as involved as a Kaplan–Meier survival curve, depending on context.

What follows is an elaboration of these aspects.

2.1 *Candidate Binary Splits*

The plethora of possible splits in 1 above—resulting from not placing any restrictions on the region A—is reduced to a computationally feasible number by constraining that:

(a) Each split depends on the value of only a *single* predictor variable. [*Note*: This restriction can be loosened; the software (CART^{TM}, California Statistical Software, 1984) permits splits on *linear* combinations of predictors.]

(b) For ordered predictors X_j, only splits resulting from questions of the form "Is $X_j \leq c$?" for $c \in \mathscr{R}^1$ are considered.

(c) For categorical predictors all possible splits into disjoint subsets of the categories are allowed.

It may appear that any reduction in number of splits resulting from the above constraints is worthless. Certainly (a) restricts us to examining predictors univariately and (b) restricts us to dividing $\mathscr{R}^1$ into two semi-infinite intervals as opposed to the multitudes of other possible break-ups. However, we are still contending with an uncountably infinite number of partitions as c ranges over $\mathscr{R}^1$. The point is that the random variable X_j takes on only a finite number of values in the sample at hand—at most n for the n cases. Hence, we have to examine only those values of c that result in a case switching "sides"—from the right semi-infinite interval to the left. So there are at most $n - 1$ splits given by $\{\text{Is } X_j \leqslant c_i?\}$ where the c_i are taken, by convention, halfway between consecutive distinct observed values of X_j.

The tree itself is grown as follows. For each node (the initial or *root* node comprises the entire sample):

1. Examine every allowable split on each predictor variable.
2. Select and execute (create two new daughter nodes) the *best* of these splits.

Steps 1 and 2 are then reapplied to each of the daughter nodes, and so on.

2.2 *Goodness-of-Split Criterion*

"Best" in 2 above is assessed in terms of the goodness-of-split criterion. Two such criteria are espoused in *CART* and are available in the associated software. These are Least Squares (§§8.3, 8.4) and Least Absolute Deviations (§8.11). Both afford a comparison based on a subadditive "between/within" decomposition, where between alludes to the homogeneity or loss measure applied to the parent node. For point of reference the definition of the least squares criterion is presented here. The obvious changes give rise to least absolute deviations (or any other between/within criterion such as is used in §6.4).

Let t designate a node of the tree. That is, t contains a subsample $\{(\vec{x}_n, y_n)\}$. Let $N(t)$ be the total number of cases in t and let

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{\vec{x}_n \in t} y_n$$

be the node response average. Then the within-node sum of squares is given by

$$\mathrm{SS}(t) = \sum_{\vec{x}_n \in t} [y_n - \bar{y}(t)]^2.$$

Now suppose a split s partitions t into left and right daughter nodes t_L and t_R. The least squares criterion is

$$\phi(s, t) = \mathrm{SS}(t) - \mathrm{SS}(t_\mathrm{L}) - \mathrm{SS}(t_\mathrm{R})$$

and the best split s^* is the split such that

$$\phi(s^*, t) = \max_{s \in \Omega} \phi(s, t)$$

where Ω is the set of all permissible splits.

A least squares regression tree is constructed by recursively splitting nodes to maximize the above ϕ function. The criterion is such that we create smaller and smaller nodes of progressively increasing homogeneity. The subadditivity mentioned above is equivalent to the nonnegativeness of ϕ; i.e., $\mathrm{SS}(t) \geqslant \mathrm{SS}(t_\mathrm{L}) + \mathrm{SS}(t_\mathrm{R})$. The same is true if we work with

least absolute deviations. It is this inequality that is responsible for the increasing (strictly nondecreasing) homogeneity.

2.3 *Determining Desirable Tree Size*

It remains unresolved as to what constitutes an appropriate sized tree. Originally [the AID program, Morgan and Sonquist (1963)] this was determined by use of stopping rules: if a node became too small or the improvement $\phi(s^*, t)$ resulting from the best split was not sufficient (to surmount some preset threshold), then the node was declared terminal. This proved unsatisfactory on account of the rigidity of the thresholds. In some instances overfitting (as a consequence of setting thresholds too small) via too large a tree would occur. In others, underfitting would result from rejection of a split (owing to its improvement not exceeding the cutoff) precluding the emergence of subsequent worthwhile splits. There is an analogy here to stepwise versus subset regression in terms of failure to capture important interactions. The problem was redressed by (1) initially growing a very large tree; (2) iteratively *pruning* this tree all the way back up (to the root node), thereby creating a nested sequence of trees; (3) selecting the best tree from this sequence using test-sample or cross-validation estimates of error.

This procedure is detailed in *CART* (Chap. 3). The means for performing the pruning in step (2) is called "minimal cost–complexity pruning"; see §3.3. This paper presents some alternatives to this when no within-node cost is available [for least squares the within-node cost is just $SS(\cdot)$]. Such alternatives are necessary because the minimal cost–complexity algorithm relies crucially on within-node cost. It is clear how the above procedure averts the deficiencies of stopping rules. Any potentially consequential splits have the opportunity to emerge, as the initial large tree can be grown so big as to possess terminal nodes that are *pure*—contain only one category in the classification context or only one response value for regression. The usage of cross-validation or test-sample estimates is intended to ensure that a realistic sized tree is produced in relation to how much noise/sampling variability is present. This aspect is discussed more fully in Section 4.

3. Two-Sample Statistic Splitting

3.1 *Motivation*

Instead of gearing our splitting criteria to optimizing within-node homogeneity, we could reward splits that resulted in large between-node separation. The magnitude of any two-sample statistic affords such a goodness-of-split measure. Such a change constitutes more than just a rephrasing of the problem. Whilst it is (empirically) the case that splitting based on two-sample t statistics with unpooled variance estimates (Welch statistics) gives results strongly resembling those obtained from least squares splitting, there is no algebraic equivalence and problems can be contrived where results are dissimilar.

The fact that in all the cases analysed, splitting using Welch statistics and splitting using least squares gave comparable results, supports the usage of two-sample statistics: given that the two techniques produce analogous results *and* least squares gives worthwhile answers, the new approach must be doing something reasonable. But why replace a proven method with one that is harder to motivate and offers no computational savings? The answer lies in the advantages provided by using two-sample *rank* statistics. These include all the conventional desiderata of ranks plus some additional benefits:

1. *Invariance under monotone transformation of the response Y.* The regression trees created by using least squares or least absolute deviations possessed such invariance only with respect to monotone transformations of the (ordered) predictors. This means, for

instance, that the optimal split is the same regardless of whether we use X_1 or $\tilde{X}_1 = g(X_1)$ for some monotone g. If the optimal split on X_1 is $X_1 \leqslant c$ then the optimal split using $\tilde{X}_1$ will be $\tilde{X}_1 \leqslant g(c)$ (§2.7). However, it is only through the use of two-sample rank statistics that best splits (predictor and cutoff) are preserved under monotone transformations h, of Y to $\tilde{Y} = h(Y)$. This is clearly a worthwhile property when there is no natural response scale in which to work (see Gordon and Olshen, 1978; Anderson, 1966). The issue is especially pertinent in the context of censored regression; the censored regression setting is further emphasized by Prentice (1978).

2. *Insensitivity to outliers in the response space.* The use of least squares, and to a lesser extent least absolute deviations, is subject to the familiar sensitivity to extreme Y observations. This, in the regression tree setting, is not necessarily a drawback, since such outliers will be isolated into their own (single case) terminal nodes. Still, the influence on overall tree topology can be distorting and the interpretation of splits leading to the isolation of the outlier can be problematic. Friedman (1979) regards the presence of outliers as weakening the least squares procedure by wasting splits.

3. *Computational feasibility.* The actual computational implementation for evaluating the multitude of competing splits is detailed in Segal's unpublished Ph.D. thesis. The easiest case is the uncensored setting, where using the Wilcoxon procedure is no more involved than using any two-sample *linear* rank statistic. The updating algorithm devised makes for an $O(n)$ procedure that is as simple as the $O(n)$ least squares algorithm. The story is not so simple when it comes to dealing with splitting based on members of the Tarone–Ware or Harrington–Fleming classes. Nevertheless, efficient algorithms can be developed with the right organization.

4. *Extension to censored response.* The principal motivation for changing the splitting criterion was to enable tree techniques to be used for survival data. Instead of using two-sample rank statistics for uncensored values as goodness-of-split criteria, analogues of such statistics that account for censoring are used, as described below.

3.2 *Censored Data Rank Statistics*

Before the merits or otherwise of the two-sample statistics used for censored response (Gehan, Prentice, Mantel–Haenszel, Tarone–Ware) are discussed, their form and computational implementation are described. The Tarone–Ware class of statistics derives from a sequence of 2×2 tables. In the survival analytic context of censored response, this sequence arises from constructing a 2×2 table for each distinct, uncensored response:

	Dead	Alive	
Population 1	a_i		n_{i1}
Population 2			
	m_{i1}		n_i

and the statistics have the following form:

$$TW = \frac{\sum_{i=1}^{k} w_i[a_i - \mathrm{E}_0(A_i)]}{[\sum_{i=1}^{k} w_i^2 \mathrm{var}_0(A_i)]^{1/2}}$$

where A_i is the random variable corresponding to number of deaths in population 1 for the ith table; w_i are constants used to weight the respective tables; the sum is over all tables, i.e., all distinct uncensored observations; the null hypothesis is that the death rates for the

two populations are equal; for fixed margins the null expectations and variances are hypergeometric:

$$E_0(A_i) = \frac{m_{i1} n_{i1}}{n_i}$$

$$\text{var}_0(A_i) = \left[\frac{m_{i1}(n_i - m_{i1})}{n_i - 1}\right]\left[\left(\frac{n_{i1}}{n_i}\right)\left(1 - \frac{n_{i1}}{n_i}\right)\right]$$

Standard specifications for the weights w_i include the following:

1. $w_i \equiv 1$ gives the Mantel–Haenszel or log-rank (Peto and Peto, 1972) statistic.
2. $w_i = n_i$ gives the Gehan (1965) statistic.
3. $w_i = n_i^{1/2}$ gives a statistic advocated by Tarone and Ware (1977).
4. $w_i = S_i^*$ gives Prentice's (1978) generalization of the Wilcoxon, where $S_i^* = \prod_{j=1}^{i} n_j/(n_j + 1)$ is almost the Kaplan–Meier survival estimate at the ith uncensored failure time.

The Gehan statistic is subject to domination by a small number of early failures (Prentice and Marek, 1979) and hence should be used selectively. Another possibility is to use the Harrington and Fleming (1982) class, which has weights $w_i = \hat{S}_i^{\rho}$ for some fixed power ρ, with $\hat{S}$ now being exactly the Kaplan–Meier survival estimate. For this to be computationally feasible it would be necessary to restrict all splits to a particular ρ value—trying any sort of optimization on ρ for individual nodes would be too involved. In practice, at least for the data sets and simulations examined, the trees emerging from using differing split statistics (apart from the Gehan) are surprisingly similar. Such a finding applied in the uncensored setting when competing members of the class of linear rank statistics were used and is also reported in *CART* in the classification-tree context.

The implementation of a splitting algorithm using the Tarone–Ware class is indicated by the following pseudo-code:

Splitting Algorithm

```
For each node
Initialize BestStat = BestPredictor = BestSplitPoint = 0
Sort response values and collect ties (compute m_i1)
Compute risk-sets n_i and the Kaplan–Meier for the node
Loop over all p predictors: X_1, X_2, . . . , X_p
  Sort the node (response and censoring indicator) with respect to ordered X_current
  Loop over all potential split points i (i.e., step along the X_current axis)
    By cycling over the ordered response values compute a_i and n_i1
    Using prespecified weights w_i compute the two-sample statistic: TwoSam
    If (| TwoSam |) > BestStat Then
      BestStat ← TwoSam
      BestPredictor ← X_current
      BestSplitPoint ← i
    End If
  End
End
```

Important but simple issues such as not splitting on tied predictor values and efficient means for sorting and ranking have not been highlighted for clarity. There is an additional user-specified parameter that serves to regulate the degree of censoring permitted in any given node. Thus, only those splits that result in daughter nodes having a ratio of uncensored to censored observations greater than the preset threshold are examined. Actual choices for this threshold should be determined in an exploratory, problem-specific manner.

4. Revised Pruning Strategies

An important difference between least squares (or least absolute deviations) splitting as outlined in Section 2 and any two-sample statistic splitting rule (§3) is that the former provides a within-node estimate of error, namely SS(t), the within-node sum of squares. Such is not the case for two-sample statistic splits, which afford only a measure of goodness of split. These measures *cannot* be decomposed to attribute a within-node error. This is consequential, since the within-node errors form a key component of the pruning algorithm developed in *CART* (Chap. 3). The algorithm, therefore, does not carry over to the present situation and, inasmuch as tree size is a fundamental issue, alternate approaches must be sought. The following section describes an attempt to inherit the *CART* algorithm and, on account of the limitations of this attack, the succeeding section details an alternative approach to pruning. This latter technique, which retains the bottom-up tactic but sacrifices cross-validation, is believed to work well.

4.1 *Inheriting the CART Algorithm*

In order to circumvent the problems posed by the absence of within-node loss, an attempt to revert back to the original splitting criteria was made. Specifically, least absolute deviations splitting was tried. Using least squares was not entertained because of the unstable nature of the mean when estimated from survival curves. The intention was to account for the censoring by using medians based on the Kaplan–Meier survival curve estimated for each node and then consider absolute deviations about these. Thus, the goodness-of-split criterion for a split at s of node t into $t_L \cup t_R$ would be

$$\phi(s,t) = \sum_{\tilde{x}_n \in t} |y_n - \nu(t)| - \sum_{\tilde{x}_n \in t_L} |y_n - \nu(t_L)| - \sum_{\tilde{x}_n \in t_R} |y_n - \nu(t_R)|,$$

where $\nu(\cdot)$ is the Kaplan–Meier median. The best split of a node t would again be that which maximized ϕ.

However, there were a variety of difficulties associated with this approach that rendered it useless:

(a) Computationally this method was very slow.

(b) The actual splits obtained using this criterion on simulated data with known structure were not convincing. The method did not uncover the important variables or split points.

(c) The hope behind resurrecting least absolute deviations splitting was the inheritance of the pruning algorithm used in standard *CART*. However, even this did not materialize. An unstated necessity for the minimal cost–complexity algorithm (§3.3) to work is that the splitting criteria be *subadditive*, i.e., $\phi(s, t) \geq 0 \ \forall s, t$. This is easily seen to hold in the uncensored case, where $\nu(t)$ above can be taken to be any sample median for the node t. But this does *not* hold for censored y's and $\nu(\cdot)$ a Kaplan–Meier median.

4.2 *Bottom-Up Approaches*

In view of the above failures, what was needed was an altogether different tack. It was decided to preserve the concept of initially growing a very large tree and subsequently pruning this. What was sacrificed was the selection of a particular tree from the generated sequence by cross-validation. Further, the minimal cost–complexity pruning algorithm itself was (by necessity) replaced with some new pruning schemata.

The loss of cross-validation as a selection mechanism was not tragic. While the method had performed well its usage had several recognized flaws. The more detracting of these include: (i) inaccuracies and instabilities of the cross-validation estimates (§8.7) and

(ii) failure of the tree selected as optimal to preclude *noisy* splits (§8.6). Of course, criticism (ii) can be levelled at any technique, but is cited here on account of such noisy splits emerging even in a highly structured situation. Indeed, the authors of *CART* equally promote user selection of the right-sized tree (§§3.4.3, 6.2). This should be done in an exploratory fashion and aided by the incorporation of subject-matter knowledge.

But for such user selection, the user must be provided with a tree sequence and hopefully one that contains good candidate trees. It was to this end that the new pruning algorithms were created. Before expounding on these it is important to reiterate what is being acquired from the *CART* approach—protection against the deficiencies of stopping rules as highlighted in Section 2.

After several strategies were tried, the following emerged as the preferred pruning algorithm:

Initially grow a very large tree.

From the bottom, step up this tree, assigning to each internal node the *maximum* split statistic contained in the subtree of which the node under consideration is the root.

Collect all these maxima and place them in increasing order.

The first pruned tree of the sequence corresponds to locating the highest node in the tree possessing the smallest maximum and removing all its descendents.

The second tree of the sequence is then obtained by reapplying this process to the first tree and so on until all that remains is the root node.

This procedure is illustrated in conjunction with the example in the next section. The associated output is also displayed. Essentially, each internal node is linked with the maximum split statistic contained in the subtree for which the node is the root. The pruning sequence is then determined by the order of these maxima. Selecting a tree from the sequence provided can be done by plotting maximal subtree split statistics against tree size and picking the tree corresponding to the characteristic "kink" in the curve; see §3.4.3 or Friedman (Technical Report 12, Department of Statistics, Stanford University, 1985).

In terms of computation time, the construction of the tree sequence is very much a secondary concern relative to the initial growing of the large tree. The building process requires the evaluation of many splits at each node, whereas for this particular pruning method, there is one very rapid ascent of the tree (to ascertain the maximum subtree split statistic for each node only simple comparisons as opposed to calculations are required), followed by subtree removal, which entails simple looping to update quantities such as number of terminal nodes. Thus, no computational burden results from the pruning algorithm.

5. Stanford Heart Transplant Data

One example for illustrating the performance of any regression technique where the response is subject to censoring is the Stanford Heart Transplant data. The parametric attacks of Miller (1976), Buckley and James (1979), and Koul, Susarla, and Van Ryzin (1981) are compared with regard to this data set by Miller and Halpern (1982). The more recent "nonparametric" treatments of Doksum and Yandell (1982), Hastie and Tibshirani (1986), and Owen (Technical Report 25, Department of Statistics, Stanford University, 1987) plus the celebrated proportional hazards model of Cox (1972) have all been tested on the Stanford data set. A fully parametric analysis on a subset of the data is presented by Aitkin, Laird, and Francis (1983).

A brief data description is now given. The response Y is $\log_{10}$ survival time, where the survival time is the time (in days) until death due to rejection of the transplant heart. There are $p = 2$ predictors: X_1, the age of the recipient, and X_2, a tissue mismatch score measuring

recipient and donor tissue compatibility. One hundred fifty-seven cases were analysed, there being a 35% censoring rate.

What has consistently emerged from the plethora of analyses is that age is the more significant predictor. For instance, Miller and Halpern (1982) exclude mismatch from additional examination, having found it to be insignificant in multiple regression analyses. However, Tanner and Wong (1983) find that patients possessing *high* mismatch and *older* ages are characterized by a distinctive hazard function. Further, the nonparametric approaches have revealed a cutoff value of roughly 50 years, in that the subpopulations so defined ($\leqslant 50$ and >50) have distinct survival characteristics.

Regression trees, using two-sample statistic splitting, were used to analyse the data. In particular, the Mantel–Haenszel statistic was used in conjunction with subtree maximal statistic pruning to produce both the tree schematic in Figure 1 and the tree sequence in Table 1. The values below the square terminal nodes in Figure 1 are Kaplan–Meier medians for that node. It is worth recording that neither the initial large tree nor the pruned tree sequence was substantially altered by using other splitting statistics from the Tarone–Ware class. In fact, the key first split was identical in the cases examined. What is immediately evident from the tree diagram is the confirmation of the previous findings. First, age clearly emerges as the more consequential predictor (but see *CART* §5.3.4 for an automated means for predictor ranking that overcomes possible *masking*— this is not an issue here since there are only two predictors). Second, the cutoff at around 50 is reflected by the value of the first split point.

However, the analysis can proceed further. A natural first summary for a terminal node when we have censored response is the estimated Kaplan–Meier survival curve $\hat{S}$. The program also provides the user with the possibility of extracting certain derived quantities,

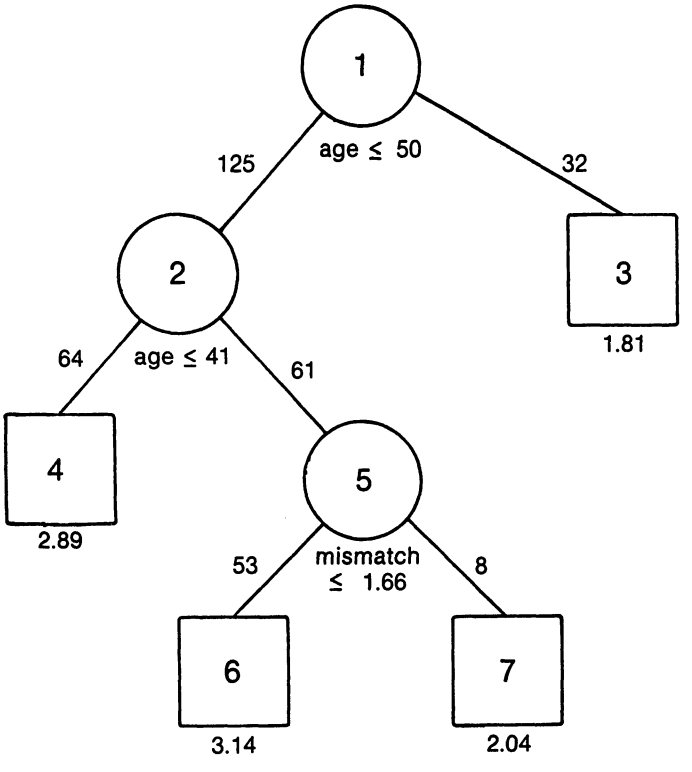

Figure 1. Tree diagram for Stanford heart transplant data.

Table 1
Pruning by subtree maximal statistics, Stanford heart transplant data

Tree number	Terminal nodes	Subtree maximal statistic
1	14	.00
2	13	.60
3	12	.99
4	11	1.17
5	10	1.25
6	9	1.79
7	6	2.43
8	4	2.48
9	2	2.79
10	1	4.91

such as the Kaplan–Meier median $\hat{S}^{-1}(.5)$ as node summaries or predictions. Figure 2 features superimposed survival curves for each of the 4 terminal nodes. The curve corresponding to node 3 in the tree schematic of Figure 2 lies appreciably below the curves for nodes 4 and 6. Node 3 contains the >50 age group. The survival prospects are noticeably worse than the bulk of the ≤50 group, contained in nodes 4 and 6, as would be expected. But, the survival characteristics for node 7 resemble those for node 3. Node 7 contains patients who are middle-aged (41–50) as opposed to young and who also have high tissue mismatch scores. Thus, it is not surprising that they possess equally poor survival prospects. It is the extraction of precisely such local interactions that makes the tree techniques so powerful. Whilst caution must be exercised in interpreting survival curves based on only 8 cases, Tanner and Wong (1983) identified a very similar local interaction.

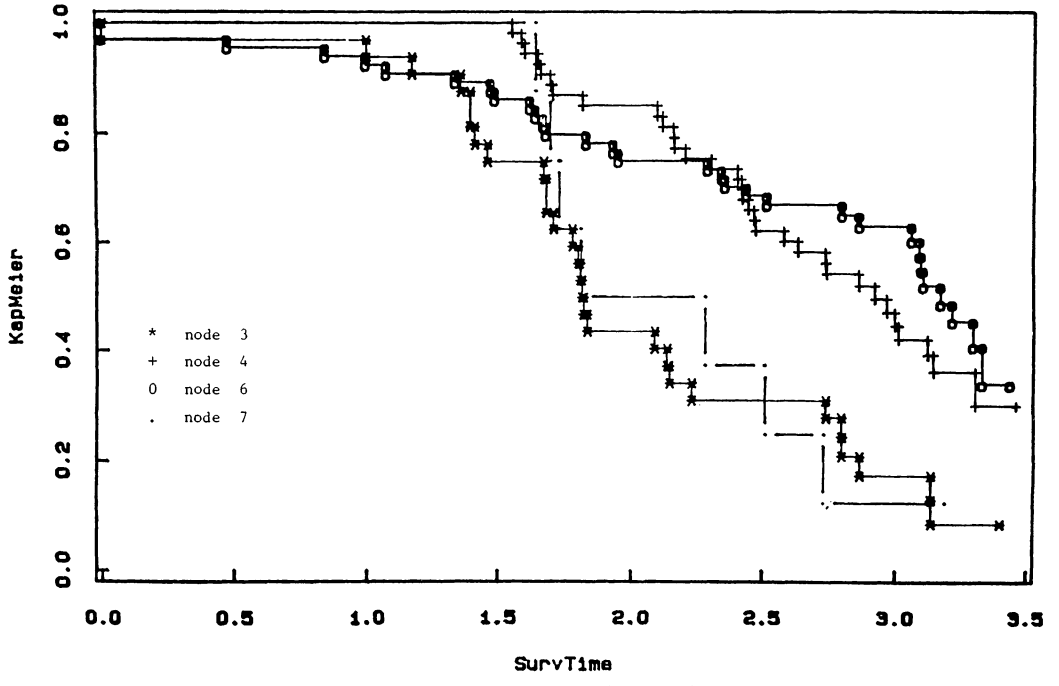

Figure 2. Kaplan–Meier curves for terminal nodes.

6. Discussion

Kalbfleisch and Prentice (1980) present a detailed analysis of mouse leukemia data to demonstrate a variety of difficulties that arise in practical problems. The particular issues, their conventional resolution, and the way in which tree structured procedures cope will form the basis of the discussion. For an actual analysis of the data using the techniques espoused here see Segal (unpublished Ph.D. thesis, Department of Statistics, Stanford University, 1986).

Kalbfleisch and Prentice (1980) formulate the practical problems encountered in terms of the following five questions, each of which is discussed in turn: (1) How should classes be formed? (2) How to deal with the multiple comparisons issue? (3) How should potential differences between Mantel–Haenszel and Gehan tests be reconciled? (4) Are sample sizes adequate for asymptotics? (5) How to cope with cases possessing missing values?

The first question pertains to Kalbfleisch and Prentice's stratification of the data. This is achieved by examining each predictor individually. For continuous predictors, division points are picked arbitrarily modulo some guidelines. These guidelines include the following recommendations: (i) three or four classes will often provide adequate resolution without unduly compromising efficiency; (ii) roughly equal sample sizes among classes should be pursued, though not at the expense of natural division points or the creation of highly unequal censoring rates.

However, there are problems associated with such rules of thumb. The consideration of predictors one at a time precludes the emergence of any interactions. The first recommendation is revealed to be inappropriate and the second to be internally inconsistent with respect to the mouse leukemia data in Segal's thesis. The regression tree approach has, as a central aim, the formation of meaningful classes. These are determined by the data themselves and hence are not subject to the vagaries of assumptions or guidelines. Interactions are readily recognized and no problems arise in dealing with variables of continuous, mixed, or discrete type. This contrasts with other nonparametric procedures based on smoothing that have difficulty dealing with mixed predictors.

The process of class formation is a precursor to testing for differences between classes. If a large number of classes are created and a large number of (nonindependent) tests performed, then the significance levels that can be attached are subject to the familiar degradation—the multiple comparisons problem. Of course, this can be accounted for by the conventional (Scheffé, Tukey, Bonferroni) methods. The problem does not exist in conjunction with regression trees, simply because no testing is performed. Whether this constitutes an asset or a liability is a contentious point. Still, more can be said on the matter. Kalbfleisch and Prentice (1980) assert that class formation on the basis of the observed mortality itself would invalidate the corresponding tests. This is true to a certain extent, yet it does not imply that, in the regression tree setting where classes *are* formed on the basis of the observed mortality data, testing is not possible. What is necessitated is that the "corresponding" tests be made to conform to the process by which the classes are constructed. Thus, for instance, in using a regression tree derived from Mantel–Haenszel splitting to perform a test on the significance of any given split (and hence of different survival curves for the two associated classes), what is required is the null distribution of the maximum of the relevant number of Mantel–Haenszel statistics. Clearly this distribution is hard to get a handle on, but the simulation studies decribed in Segal's thesis can be used instead.

With regard to the third question on reconciling differences, should they arise, between the Mantel–Haenszel and Gehan tests for equality of survival curves from two classes, Kalbfleisch and Prentice (1980) advise using some intermediary weighting of the tests, i.e., some member of the Tarone–Ware class. This is completely concordant with the splitting

strategy used in the construction of regression trees, whereby any member of the Tarone–Ware or Harrington–Fleming classes can be used as the splitting criterion.

Again, since no formal testing is undertaken in the regression tree approach, the fourth question concerning asymptotics is somewhat moot. The recommendation of Kalbfleisch and Prentice, for situations where sample sizes are perceived to be too small to warrant recourse to asymptotic results, is to perform testing by simulating the actual null distribution of the statistic. This is precisely the strategy proposed in conjunction with regression tree methods. If censoring is independent of the predictors, an alternative simulation tactic would be to develop a permutation test. This would involve permuting the response values and corresponding censoring indicators over the cases and then recomputing the tree.

The only asymptotic issue to be examined in the context of tree schemata is consistency; see *CART* (Chap. 12) and the sequence of papers by Gordon and Olshen (1978, 1980, 1984). Under regularity conditions that do *not* depend on the particular splitting criteria or pruning algorithm used, consistency results are obtained for both the classification and regression problems. The regularity conditions include a growth condition on the amount of mass in each member of the partition and the requirement that the diameter of every member go to 0 in probability. For censored response data, identifiability issues arise as indicated in Gordon and Olshen (1985), yet since there is no reliance on the particular tree construction methods used, the consistency results carry over immediately to the two-sample statistic schemes developed here.

The final question concerns how to cope with missing values. These can constitute a consequential portion of the data in medical/biological studies and hence efficient information extraction from (as opposed to the discarding of) such cases is an important issue. The manner in which this is achieved by tree schemata is detailed in *CART* (Chap. 5). For an illuminating illustration of how distorted results can occur using conventional regression practices in the presence of missing data, and how tree methods overcome such problems, see Bloch and Segal (Technical Report 108, Department of Statistics, Stanford University, 1985).

The overall conclusion then, is that the regression tree methodology developed deals very well with the five problems posed as being of practical consequence.

Acknowledgements

The author wishes to thank Professor Jerome Friedman and the two referees for many helpful comments.

Résumé

On étend la méthodologie des arbres de régression aux variables de réponse censurées à droite en remplaçant les règles habituelles de coupure par des règles basées sur les classes de Tarone–Ware ou Harrington–Fleming de statistiques pour deux échantillons. On donne aussi de nouvelles stratégies d'élagage permettant de déterminer la taille adéquate de l'arbre. On développe les propriétés de cette approche, et on discute des comparaisons avec les procédures existantes, en termes de problèmes pratiques. Une illustration des performances réelles de la technique est présentée.

References

Aitkin, M., Laird, N., and Francis, B. (1983). A reanalysis of the Stanford Heart Transplant Data. *Journal of the American Statistical Association* **78,** 264–274.

Anderson, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. In *Multivariate Analysis*, P. R. Krishnaiah (ed.), 5–27. Orlando, Florida: Academic Press.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80,** 580–598.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Belmont, California: Wadsworth.

Buckley, J. and James, I. R. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Ciampi, A., Chang, C.-H., Hogg, S., and McKinney, S. (1987). Recursive partition: A versatile method for exploratory data analysis in biostatistics. In *Proceedings from Joshi Festschrift,* G. Umphrey (ed.), 23–50. Amsterdam: North-Holland.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–202.

Doksum, K. A. and Yandell, B. S. (1982). Properties of regression estimates based on censored survival data. In *Festschrift for Erich L. Lehmann,* P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr. (eds), 140–156. Berkeley: University of California Press.

Friedman, J. H. (1979). A tree-structured approach to nonparametric multiple regression. In *Smoothing Techniques for Curve Estimation,* T. Gasser and M. Rosenblatt (eds), 5–22. Berlin: Springer-Verlag.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76,** 817–823.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52,** 203–223.

Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., et al. (1982). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *New England Journal of Medicine* **307,** 588–596.

Gordon, L. and Olshen, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Annals of Statistics* **6,** 515–533.

Gordon, L. and Olshen, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis* **10,** 611–627.

Gordon, L. and Olshen, R. A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis* **15,** 147–163.

Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* **69,** 1065–1069.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science* **1,** 297–318.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics* **9,** 1276–1288.

Miller, R. G., Jr. (1976). Least squares regression with censored data. *Biometrika* **63,** 449–464.

Miller, R. G., Jr. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69,** 521–531.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58,** 415–434.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank-invariant test procedures. *Journal of the Royal Statistical Society, Series A* **135,** 185–198.

Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika* **65,** 167–179.

Prentice, R. L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35,** 861–867.

Stone, C. J. (1977). Consistent nonparametric regression (with Discussion). *Annals of Statistics* **5,** 595–645.

Tanner, M. A. and Wong, W. H. (1983). Discussion of: A reanalysis of the Stanford Heart Data by Aitkin, Laird, and Francis. *Journal of the American Statistical Association* **78,** 286–287.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64,** 156–160.

Received September 1986; *revised June* 1987.